Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

A Comprehensive Investigation into the Accuracy of Soft Computing Tools for Downscaling Precipitation Data Extracted from Climate Models

Shalaka Shah [1]; Shreenivas N. Londhe [2,*]

1. Research Scholar, Department of Civil Engineering, Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
2. Professor, Department of Civil Engineering, Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
* Corresponding author: *shreenivas.londhe@viit.ac.in*

https://doi.org/10.22115/SCCE.2024.403571.1674

ARTICLE INFO

Article history:
Received: 18 July 2023
Revised: 17 April 2024
Accepted: 10 May 2024

Keywords:
Climate change;
Support vector regression;
Model trees;
Artificial neural networks;
Multi-gene genetic programming;
Rainfall.

ABSTRACT

This study will provide the audience with an understanding of the capabilities of soft tools such as Artificial Neural Networks (ANN), Support Vector Regression (SVR), Model Trees (MT), and Multi-Gene Genetic Programming (MGGP) as a statistical downscaling tool. Many projects are underway around the world to downscale the data from Global Climate Models (GCM). The majority of the statistical tools have a lengthy downscaling pipeline to follow. To improve its accuracy, the GCM data is re-gridded according to the grid points of the observed data, standardized, and, sometimes, bias-removal is required. The current work suggests that future precipitation can be predicted by using precipitation data from the nearest four grid points as input to soft tools and observed precipitation as output. This research aims to estimate precipitation trends in the near future (2021-2050), using 5 GCMs, for Pune, in the state of Maharashtra, India. The findings indicate that each one of the soft tools can model the precipitation with excellent accuracy as compared to the traditional method of Distribution Based Scaling (DBS). The results show that ANN models appear to give the best results, followed by MT, then MGGP, and finally SVR. This work is one of a kind in that it provides insights into the changing monsoon season in Pune. The anticipated average precipitation levels depict a rise of 300–500% in January, along with increases of 200-300% in February and March, and a 100-150% increase for April and December. In contrast, rainfall appears to be decreasing by 20-30% between June and September.

How to cite this article: Shah S, Londhe SN. A Comprehensive investigation into the accuracy of soft computing tools for downscaling precipitation data extracted from climate models. J Soft Comput Civ Eng 2025;9(1):158–178. https://doi.org/10.22115/scce.2024.403571.1674

2588-2872/ © 2025 The Authors. Published by Pouyan Press.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

The southwest monsoon precipitation (June to September) dominates the Indian climate. However, general observation over the last few years indicates that there is a variation in the monsoon season time frame, particularly in the city of Pune, India. [1,2] These changes bring with them a slew of new challenges for water resource planning and management.

Climate data is sourced from General Circulation Models (GCMs) worldwide. These sophisticated tools are adept at projecting future climate variations [3,4], designed to replicate atmospheric, oceanic, and terrestrial processes. GCMs serve as the foundation for computer programs that simulate atmospheric conditions by applying equations like the Navier-Stokes equations on a rotating Earth and incorporating thermodynamic terms for different energy sources. These models are run for various warming scenarios known as Representative Concentration Pathways (RCPs), offering a spectrum of potential futures for atmospheric composition evolution [3].

General Circulation Models (GCMs) employ a three-dimensional grid spanning the globe, typically with a horizontal resolution ranging from 250 to 600 km. These models consist of 10 to 20 vertical layers in the atmosphere and often around 30 layers in the oceans. However, due to their relatively coarse spatial resolutions, GCMs may struggle to accurately represent certain sub-grid scale features at smaller scales. Consequently, values derived from different grid points for any physical processes need to be downscaled to align with the actual regional or local environmental conditions [3,4].

The process of downscaling GCMs at local and regional levels has prompted the development of models where a finely computed grid within a limited area is nested within the broader grid of a GCM. This technique is known as dynamic downscaling. However, dynamic downscaling has a significant drawback that limits its application in climate change impact studies: its intricate design and high computational demands. Additionally, it lacks flexibility, as expanding the area of study or shifting to a slightly different region necessitates repeating the entire experiment. Another downscaling approach, called statistical downscaling, involves establishing empirical relationships that transform large-scale GCM features (Predictors) into regional-scale variables (Predictands), such as precipitation and streamflow [5].

Numerous investigations in India have examined the effects of climate change on rainfall. Studies referenced as [6–10] have indicated a weakening trend in southwest monsoons due to climate change. Conversely, research by [11–13] suggests the potential for increased monsoon rainfall across South Asia attributable to global warming. Furthermore, [13,14] assessed the impact of climate change in India by analyzing the PRECIS climate model's simulation for the period 1961-1990. Their findings revealed a rise in extreme rainfall along the west coast and western regions of central India.

The study done by [15] examined the impact of climate change on precipitation in Mumbai. Their findings suggest a projected increase in future rainfall intensity from 2010 to 2099. Over a 30-year period, they anticipate an average increase of 15-20% in maximum rainfall, and over a 90-year period, this increase escalates to 30-45%. Additionally, their research highlights a shift in

the southwest monsoon season and a delayed onset of the monsoon period. Employing a statistical downscaling technique known as the Distribution-based Scaling technique (DBS), they scaled the GCM data. The results of this scaling approach, aimed at correcting GCM projections to the local scale, are deemed potentially valuable for impact studies.

In the all the aforementioned studies one or the other statistical downscaling technique is used which requires pre-processing (in the form of re-gridding when the grid points of the GCM differed from the grid points of the observed data) and bias removal which makes the methods computationally intensive. If the downscaling can be achieved by employing techniques which can avoid the pre-processing and bias removal without sacrificing accuracy, such techniques should always be welcomed. The soft computing techniques which work on data rather than physics of the process can serve the purpose due to their model free nature and learning capabilities necessitates a comprehensive a study to compare performances of few such techniques with each other as well as with one traditional technique which is the motivation behind this work. In the present work soft computing techniques of Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Model Trees (MT) and Multi-gene Genetic Programming (MGGP) are employed to predict rainfall in Pune, the 8th largest of India, along with traditional technique of Distributed based scaling technique (DBS)

Soft computing tools like Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Model Trees (MT) have previously been used as statistical downscaling tools. The performance of three downscaling methods, namely conditional random field (CRF), K-nearest neighbor (KNN), and support vector machine (SVM) was compared by [16], in downscaling precipitation in the monsoon-prone Punjab region of India. These models were used to downscale daily monsoon precipitation (June to September) at six locations in Punjab. As per their results, the CRF and KNN models outperform the SVM model in reproducing most daily rainfall statistics. The research done by [17] investigated the projected future monthly rainfall in Haryana using CGCM3 outputs for an A2 emission scenario. They utilized a gridded rainfall dataset from the National Climate Centre, IMD Pune, as observed data. A relationship was established between predictand and predictors from the NCEP/NCAR Reanalysis, which was then utilized to downscale rainfall for future periods. Comparing statistical performance parameters such as correlation coefficient and root mean square error, the M5P Model demonstrated superior performance over SVR. Additionally, [18] employed ANN to statistically downscale global climate models (GCMs) at meteorological sites in Bangkok, Thailand, during the rainy season. The research employed a feed-forward backpropagation network, utilizing extensive predictor variables obtained from both ERA-interim reanalysis data and current or projected GCM data. Initially, predictors were selected from various grid boxes surrounding the Bangkok area and then filtered through principal component analysis (PCA) to identify the most highly correlated predictors for ANN training. The downscaled results from the reanalyzed current climate exhibited favorable alignment with station precipitation data. Notably, the ultimate downscaled outcomes for four GCMs indicate a rising pattern in precipitation during the rainy season over Bangkok by the conclusion of the twenty-first century.

In the work done by [19], they developed multiple machine learning (ML) downscaling models utilizing a Bayesian model average (BMA) approach to downscale precipitation simulations

from 8 Coupled Model Intercomparison Project Phase 5 (CMIP5) models. Model output statistics (MOS) were employed for the upper Han River basin spanning from 1961 to 2005. The application of these downscaled models notably enhanced station simulation performance. Their findings indicated that Support Vector Machine for Regression (SVR) surpassed Multilayer Perceptron (MLP) and Random Forest (RF) models. The downscaling outcomes derived from the BMA ensemble simulation combined with SVR models demonstrated the most favorable performance overall. In a separate study, [20] investigated three statistical learning algorithms—Support Vector Machine (SVM), Random Forest Regression (RF), and Gradient Boosting Regressor (GBR)—to propose an efficient downscaling approach aimed at generating high spatial resolution precipitation data. Their methodology involved producing high-resolution precipitation data from daily station-level precipitation records and longitude/latitude/altitude data. They conducted a downscaling analysis utilizing daily observed precipitation data from 34 monitoring sites in Bangladesh to showcase the efficiency and accuracy of their models compared to traditional Multiple Linear Regression (MLR) downscaling models. Among the four downscaling models evaluated, the GBR-based approach yielded the most favorable performance.

The key novelty of this study is its methodology, which involves using precipitation data from the four closest grid points surrounding the target area as input variables for training soft computing tools. These tools include Artificial Neural Network (ANN), Support Vector Regression (SVR), Model Tree (MT), and Multi-Gene Genetic Programming (MGGP), with observed rainfall serving as the output.

While performing their primary task of downscaling precipitation data, these 4*1 models will automatically overcome the challenges of standardization, bias removal, and regridding. Secondly, this study is a one-of-a-kind comprehensive study of soft computing tools as downscaling techniques. According to the authors' knowledge, Multigene Genetic Programming has never been implemented as a statistical downscaling tool, which adds to the work's novelty. Thirdly, in the majority of studies where precipitation is modeled using causal parameters, linear measures like the coefficient of correlation are commonly employed for predictor selection. However, the present study introduces Average Mutual Information (AMI) as an input selection measure, a novel approach not previously utilized in climate change research. Fourthly, the study stands out for its detailed examination of Artificial Neural Network (ANN) models. Specifically, a four-layered neural network was employed, along with nine different training algorithms, a level of granularity rarely observed in prior studies. Fifth, to ensure model stability, a comprehensive decade-wise analysis of Root Mean Square Error (RMSE) was conducted and graphically presented for the previous 116 years, maintaining a constant vigil on result accuracy. Sixth, this research underscores the capability of ANNs to accurately predict precipitation even when some causal parameters are absent during training, affirming the network's robustness in handling missing data.

Moreover, while the literature review primarily focuses on regional-level studies, this research addresses the need for localized analysis by examining the erratic rainfall patterns and monsoon season shifts in Pune, India. Remarkably, no comprehensive analysis of future rainfall trends for Pune has been previously conducted, making this study a significant contribution to the field.

The next section throws light on the study area and data used in this research, followed by a concise explanation of the working of the various soft computing tools used. Section 4 describes the methodology and model formulation. The results and related discussion are presented in section 5. Section 6 provides the details of the possible future precipitation trends in Pune followed by concluding remarks in section 7.

2. Study area and data

The research is centered on Pune, a city situated in the state of Maharashtra, India. Pune ranks as the eighth largest city in the country, boasting a population exceeding 5 million people. Renowned as a significant educational, information technology, and automobile hub, Pune holds strategic importance. Figure 1 illustrates the geographical layout of Pune.

The observed rainfall data for the study was generously provided by the India Meteorological Department (IMD), Pune, accessible at their website (https://www.imdpune.gov.in/). This dataset spans from 1901 to 2016, covering 116 years without any gaps in cumulative monthly precipitation values.

In addition to the observed data, various climate parameters including rainfall, surface humidity, maximum and minimum temperatures, pressure at mean sea level, and 10 m wind speed were sourced from the Copernicus Climate Change Service website (https://cds.climate.copernicus.eu/). These data were derived from Global Climate Models (GCMs) and utilized in the study.

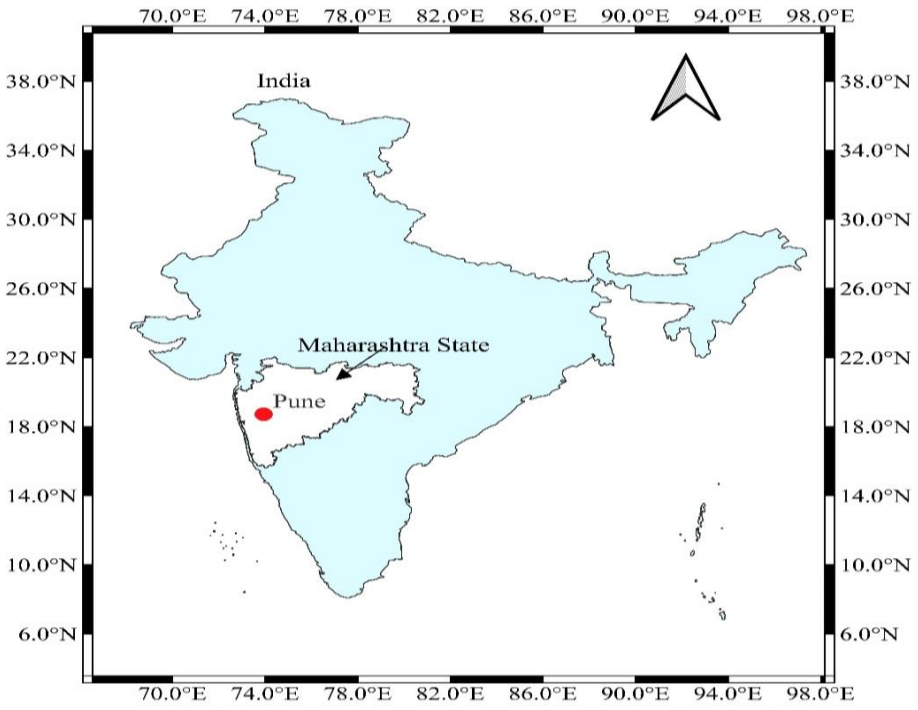

Fig. 1. Study Area: Pune, located on the map of India at coordinates 18.5204° N latitude and 73.8567° E longitude.

3. Soft computing tools

Soft computing techniques are a key instrument in the applications of artificial intelligence that are emerging today. For decades, researchers have been studying the hydrological applications of these soft tools. The current study is one such example, in which four soft computing techniques are used to downscale and predict future rainfall using data extracted from five global climate models (GCMs). Artificial Neural Network (ANN), Support Vector Regression (SVR), M5 Model Tree (MT), and Multi-Gene Genetic Programming (MGGP) are the soft computing techniques used in this study.

Biological neural networks are the inspiration for Artificial Neural Networks. An Artificial Neural Network, akin to its biological counterpart, continuously learns and enhances its understanding of the environment by processing experiences. One can use neural networks to better understand the relationships between complex data structures. They can predict the behavior of complex structures using their trained weights and biases (knowledge).

Support Vector Regression (SVR) aims to approximate target values by finding a function that minimizes deviation from the true value while maintaining a flat profile to prevent overfitting. SVR tackles nonlinear regression problems by mapping low-dimensional input data into a high-dimensional feature space through kernel functions.

Model Tree (MT) resolves complex issues by decomposing them into simpler components and integrating their solutions. In MT, the splitting process resembles that of a decision tree, with leaves containing linear regression functions for predicting continuous numerical attributes instead of class labels.

Genetic Programming (GP) is a biologically inspired approach in which computer programs evolve to fulfill a task, typically represented by tree structures. Multi-Gene Genetic Programming (MGGP) introduces a concept of multigene individuals, where each gene comprises a GP tree. These genes are incrementally acquired to enhance fitness, with each gene weighted linearly and combined to form a pseudo-linear model capable of capturing nonlinear behavior.

For detailed information on Artificial Neural Networks (ANN), SVR, MT, and MGGP, readers are directed to [21–24] respectively. Despite operating on different principles, this study illustrates that these soft computing tools outperform the traditional Distribution-based scaling method proposed by [5] in downsizing GCM data. This underscores the adaptability advantage of soft tools over conventional hard tools.

4. Methodology

The historical data for all parameters from five GCMs, spanning from 1901 to 2016 and future projections covering the period from 2021 to 2050 were downloaded and extracted for analysis. Data from five Global Climate Models (GCMs), namely IPSL_CMA5, HADGEM2_CC, NorEMS1_1, CNRM_CM5, and GFDL_CM3, obtained from the Coupled Model Intercomparison Project 5 (CMIP5), were utilized in this study. These GCMs are noted for their

relatively higher accuracy in simulating climate conditions over the Indian subcontinent and have been widely employed by researchers, including those referenced in [5]. Table 1 contains information about these GCMs.

Table 1
Details of GCMs used for the current study.

GCM	CNRM_CM5	HadGEM2_CC	GFDL_CM3	IPSL_CM5A_MR	NorESMI_I
Institution	Centre National de Recherches Meteorologiques, France	Met Office Hadley Centre, UK	Geophysical Fluid DynamicLaboratory, USA	Institute Pierre-Simon Laplace, France	Norwegian Climate Centre, Norway
Grid Size (Lat * Long)	1.4008*1.4063	1.25*1.875	2*2.5	1.2676*2.5	1.8947*2.5

The current work is divided into 2 phases, Phase 1 and Phase 2. The precipitation occurring at the four nearest grid points surrounding Pune was used as input in Phase 1 of the work, and the output was the observed rainfall obtained from IMD, Pune. The models were trained and tested using ANN, MT, SVR, and MGGP. DBS was also used to downscale the rainfall. The detailed working of DBS will not be covered in this paper due to the paucity of space. Readers are referred to [9] to gain a better understanding of the DBS technique. Phase 2 was developed to overcome the drawbacks of Phase 1. The detailed need and methodology of Phase 2 has been covered later in this paper.

To evaluate the accuracy of the models, various statistical parameters and error measures were employed. The mean, standard deviation, and maximum values derived from the simulated models were compared with those from the observed dataset. Four error measures were utilized in this study. Among them were two absolute parameters, namely Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), along with two dimensionless parameters, namely Coefficient of Efficiency (CE) and Coefficient of Correlation (r). Additionally, due to the importance of precision in peak prediction, another relative error measure, Percentage Error in Peak (PEP), was incorporated. For a comprehensive understanding of all error measures, readers are directed to [25].

During the training and testing of the ANN models, it was discovered that a three-layered feed-forward back-propagation network was unable to break the complexity of the problem at hand, so an additional hidden layer was added, resulting in a four-layered FFBP architecture (not shown here). The IPSL_CMA5_MR model was trained and tested using the 9 different training algorithms namely (i) gradient descent with momentum (GDM), (ii) resilient backpropagation (RP), (iii) conjugate gradient Fletcher–Reeves update (CGF), (iv) conjugate gradient Polak–Ribiere update (CGP), (v) Powell Beale restarts (CGB), (vi) scaled conjugate gradient (SCG), (vii) Broyden, Fletcher, Goldfarb, and Shanno update (BFG), (viii) one step secant algorithm (OSS) and (ix) Levenberg–Marquardt algorithm (LM) recommended by [26] to select the best amongst them. Readers are referred to [27] for details of these algorithms. Table 2 displays the performance accuracy of the nine different algorithms. It shows that when the values of the coefficient of correlation(r), RMSE, PEP, and CE are used for comparison, LM performs significantly better than the other algorithms. As a result, the LM algorithm was used to train the rainfall modes for all 5 GCMs. Support Vector Regression and Model Tree models were trained

and tested using the Waikato Environment for Knowledge Analysis (WEKA) Software, version 3.8.6 [28]. For training the SVR models, the SMORegressor function has the following specifications:

1. Algorithm – RegOptimizer: RegSMOImproved
2. Kernel – The polynomial kernel

For training the MT models, the M5P classifier was chosen. It implements the M5P model tree algorithm. MATLAB v R2013a was used for training and testing the ANN and MGGP models. Extensive trial and error went in to calibrate the population size, maximum number of generations, generations elapsed, maximum number of genes, maximum tree depth, and maximum total nodes while training the MGGP models in MATLAB 2013a. The optimal combination of the aforementioned variables was determined. After the model has been trained, it provides us with an equation we may use to directly predict future rainfall amounts. One of the main benefits of MGGP is that, as the optimum gene combination is selected, it automatically identifies the inputs that have little or no bearing on the outcome and removes them from the solution. As a result, MGGP handles feature scaling, a crucial operation, by default.

Table 2
Comparison of performance of 9 different ANN algorithms.

GCM	Error Measure	GDM	RP	CGF	CGP	CGB	SCG	BFG	OSS	LM
IPSL - CMA5_MR	r	0.56	0.62	0.63	0.65	0.65	0.65	0.66	0.64	0.73
	RMSE (mm)	99.34	93.66	91.93	90.25	90.28	90.06	89.30	91.44	81.13
	CE	0.30	0.38	0.40	0.42	0.42	0.42	0.43	0.41	0.53
	PEP (%)	-49.01	-62.24	-55.92	-64.65	-64.97	-65.70	-69.62	-57.82	-35.36
HadGEM2 _CC	r	0.59	0.74	0.71	0.64	0.64	0.74	0.75	0.73	0.74
	RMSE (mm)	99.95	80.08	87.14	91.76	91.79	80.93	79.58	81.87	84.06
	CE	0.29	0.54	0.46	0.40	0.40	0.53	0.55	0.52	0.50
	PEP(%)	-55.29	-57.85	-52.60	-65.53	-65.45	-57.01	-56.83	-53.13	-43.81
NorESMI_I	r	0.73	0.76	0.70	0.76	0.76	0.76	0.76	0.75	0.76
	RMSE (mm)	100.99	78.67	85.83	78.38	78.48	78.45	78.33	79.56	78.20
	CE	0.28	0.56	0.48	0.56	0.56	0.56	0.56	0.55	0.57
	PEP (%)	-71.72	-57.72	-69.93	-54.36	-55.20	-54.65	-53.40	-60.03	-52.35
CNRM_C M5	r	0.65	0.73	0.74	0.74	0.74	0.74	0.74	0.74	0.75
	RMSE (mm)	96.24	81.30	79.85	79.98	79.87	79.88	79.88	80.19	77.97
	CE	0.34	0.53	0.55	0.55	0.55	0.55	0.55	0.54	0.57
	PEP (%)	-53.22	-65.32	-60.72	-61.85	-60.86	-60.36	-61.01	-53.44	-49.45
GFDL_CM 3	r	0.73	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74
	RMSE (mm)	81.62	81.04	81.05	81.05	81.69	81.02	81.04	81.04	81.04
	CE	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
	PEP(%)	-47.35	-58.01	-42.27	-41.57	-47.50	-42.19	-44.44	-41.15	-44.58

5. Results and discussion

The results of Phase 1 are shown in Table 3 and Figure 2. Table 3 shows that for the IPSL_CMA5 and HADGEM2_CC models, ANN outperforms the other techniques significantly with low RMSE, MAE and high CE ($RMSE_{ANN}$=81.13mm, MAE_{ANN}=50.93mm, CE_{ANN}=0.53) ($RMSE_{ANN}$=84.06mm, MAE_{ANN}=49.7mm, CE_{ANN}=0.50). In terms of NorESMI_1, MT appears to be the best amongst all ($RMSE_{MT}$=74.98mm, MAE_{MT}=48.00mm, CE_{MT}=0.60). The soft

computing tools, appear to perform equally well for the CNRM_CM5 and GFDL_CM3 models as well. In terms of accuracy, soft computing techniques outperform DBS for all the five GCMs.

The typical scatter plots depicted in Figure 2 support the above findings. These scatter diagrams depict the IPSL_CMA5 results for all the four soft computing techniques and DBS. The scatter plots for DBS and SVR have a high proportion of underestimated values, whereas the MT and MGGP plots consistently over-predict the results. ANN, on the other hand, appears to have a fairly balanced scatter.

Even though ANN appears to have an advantage over the other statistical downscaling techniques, the values of PEP in all models show that it still has a disadvantage in peak prediction. This can be attributed to using the rainfall values only and not the values of causative parameters which influence the rainfall.

Phase 2 methodology was created to address this shortcoming. To improve the accuracy of the Phase 1 models, causative parameters were added as inputs. Studies were conducted to focus on finalizing the inputs for cause-effect modeling using soft computing tools [29–32]. In previous studies, significant influencing parameters were typically selected using linear methods like coefficients of correlation, determination coefficients, and adjusted determination coefficients from a wide array of predictor variables. However, these linear tools may overlook the nonlinear relationships between independent and dependent variables. To address this limitation, the present study suggests employing Average Mutual Information (AMI) instead of correlation analysis to determine input variables. AMI quantifies the nonlinear relationship between predictor and predictand variables. Notably, according to the referenced literature, no prior climate change studies have utilized AMI as an input selection tool. For further information on AMI, readers are encouraged to consult [33].

Table 3

Phase 1 results (all values are in mm).

GCM	Error Measure	DBS	ANN	SVR	MT	MGGP
IPSL_CM5	RMSE (mm)	125.23	81.13	100.89	85.41	87.54
	MAE (mm)	74.10	50.93	57.04	53.83	59.49
	CE	-0.11	0.53	0.28	0.48	0.46
	PEP (%)	15.09	23.71	-23.36	-39.22	-55.06
NOR_ESMI	RMSE (mm)	99.98	78.20	83.94	74.98	75.07
	MAE (mm)	59.67	51.04	48.10	48.00	50.37
	CE	0.29	0.57	0.50	0.60	0.60
	PEP (%)	8.58	-52.35	-29.93	-46.39	-47.80
CNRM_CM5	RMSE (mm)	105.06	77.97	86.33	78.65	77.81
	MAE (mm)	63.11	49.63	48.86	49.07	49.30
	CE	0.21	0.57	0.47	0.56	0.57
	PEP (%)	31.82	-49.45	-17.37	-37.07	-44.63
GFDL	RMSE (mm)	85.93	81.03	80.14	77.03	76.79
	MAE (mm)	52.00	51.34	47.29	47.44	48.17
	CE	0.47	0.53	0.54	0.58	0.58
	PEP (%)	-21.59	-44.58	-44.62	-45.78	-40.02
HAD	RMSE (mm)	129.97	84.06	112.07	87.01	92.45
	MAE (mm)	69.32	49.70	64.65	58.62	61.35
	CE	-0.20	0.50	0.11	0.46	0.39
	PEP (%)	44.24	-43.80	-21.21	-54.37	-50.09

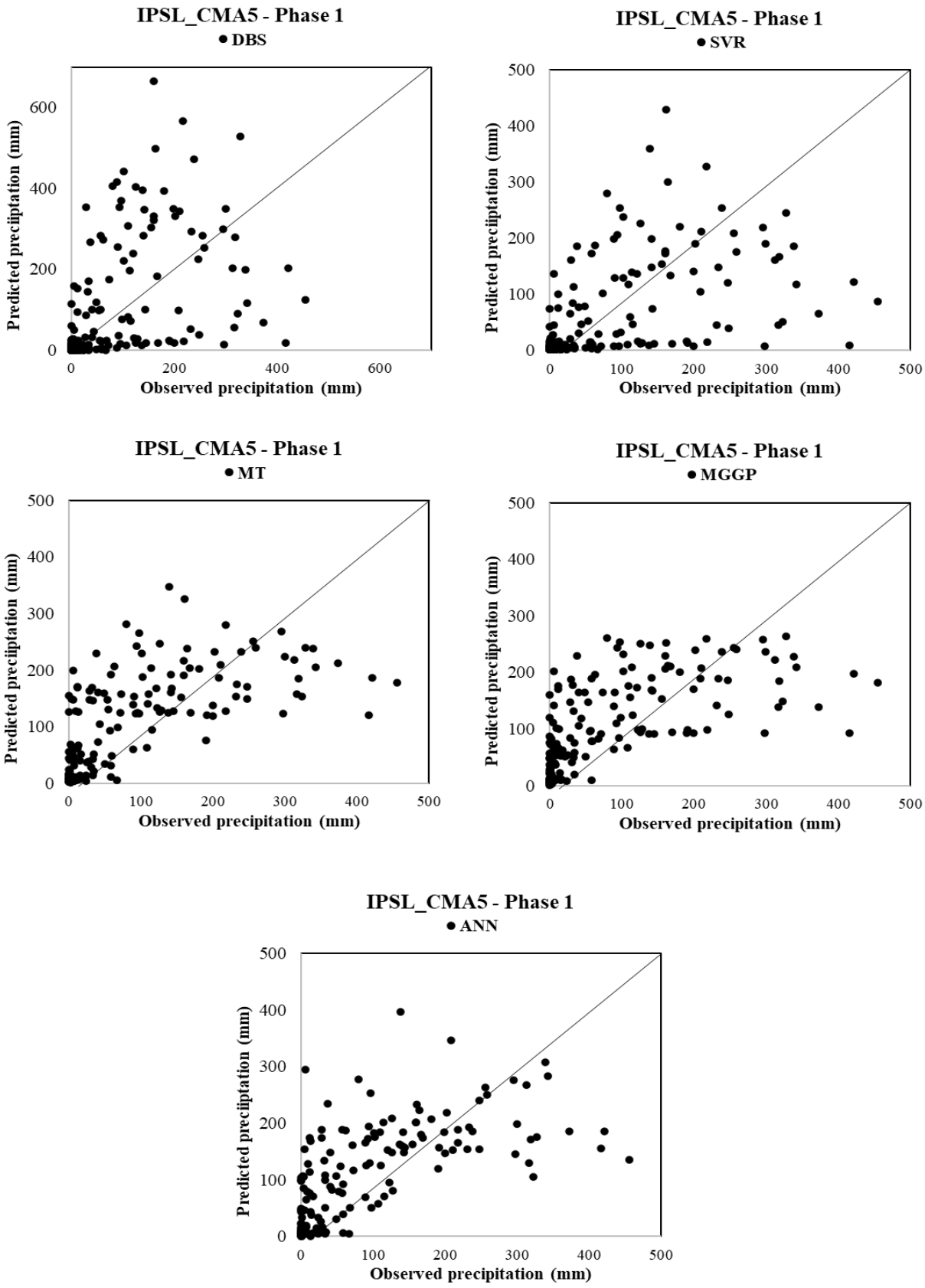

Fig. 2. (a to e). Scatter plots (IPSL_CMA5 – Phase 1).

For the current study, five causative parameters believed to influence both the magnitude and frequency of rainfall were selected, as outlined in [34]. These parameters include daily minimum and maximum temperature, pressure at sea level, 10m wind speed, and humidity.

Data from the Copernicus Climate Change Service website was utilized to acquire GCM data pertaining to minimum and maximum temperature, surface humidity, 10-meter wind speed, and pressure at mean sea level. Information for all parameters from five distinct GCMs was downloaded and extracted for both historical (1901–2016) and future (2021–2050, RCP 4.5) time frames. Subsequently, the Average Mutual Information (AMI) between the five causative parameters and observed rainfall was calculated individually for each GCM. In regions where daily temperature variation is minimal, both the minimum and maximum temperatures may not be required to train the model.

As an alternative approach, the average temperature for each day was considered as an additional causative parameter in addition to the minimum and maximum temperatures. Consequently, the average temperature was included as an input variable in the current study alongside the existing parameters. Table 4 presents the Average Mutual Information (AMI) values corresponding to each of the five GCMs. Leveraging the AMI values from Table 4, several input-output models were constructed. Table 4 shows us that in the case of IPSL_CMA5_MR, precipitation has the highest AMI value with humidity (0.50) hence model 1A consisted of humidity as an input and precipitation as an output. AMI value of precipitation with minimum temperature (0.35) ranks 2^{nd} hence model 2A has humidity and minimum temperature as inputs. In a similar manner models 3A to 5A were formulated based on the AMI values. Model 6A uses the average temperature in place of the maximum and minimum temperatures. To increase the degrees of freedom and give the soft tools additional learning information, precipitation was as well added as an input to model 7A. Details of the IPSL_CMA5_MR models can be seen in Table 5. Similarly, the models were also formulated for the remaining 4 GCMs. Models 1B-6B were developed for NorESM1_1 as wind speed was missing for this GCM. Models 1C-7C were developed for CNRM_CM5, 1D-7D for GFDL_CM3 and 1E-7E for HadGEM2_CC. The models were then trained and tested using all four soft computing tools mentioned above, namely ANN, SVR, MT, and MGGP. Table 6 shows the typical architecture of the models developed using soft computing tools (Only Model 7A architecture is shown due to paucity of space).

Table 7 depicts a typical result table, and Figure 3 depicts typical scatter plots to compare all methods.

When the results for phase 2 are studied, it can be observed that the RMSE values for ANN is consistently better as compared to the other techniques. Also, as one proceeds from model 1A to 7A, the $RMSE_{ANN}$ decreases from 70.41mm to 60.63mm. The same can be observed even for MAE_{ANN} (42.71mm to 39.93mm), CE_{ANN} (increases from 0.64 to 0.74), and even PEP_{ANN} (-52.86% to 2.63%). The negative sign in the case of PEP depicts under prediction of peaks. These results also show that by adding causative parameters to the soft tools, the peak prediction seems to have improved significantly.

Table 4
AMI Values for Model Formulation (all values are in mm).

Contributing Parameter		Max Temp	Min Temp	PSL	Humidity	Wind Speed	Avg Temp
AMI values	IPSL_CMA5	0.12	0.35	0.29	0.5	0.34	0.21
	NorESMI_1	0.16	0.32	0.39	0.49	xx	0.25
	CNRM_CM5	0.14	0.33	0.28	0.52	0.28	0.15
	GFDL_CM3	0.1	0.27	0.3	0.51	0.34	0.24
	HADGEM2_CC	0.13	0.32	0.33	0.51	0.4	0.15

Table 5
Typical model formulation for IPSL_ CMA5.

Model Name	Input	Output	Input Size: Output Size
1A	Humidity	Obs Precipitation	4:1
2A	Humidity, Min Temp	Obs Precipitation	8:1
3A	Humidity, Min Temp, Wind Speed	Obs Precipitation	12:1
4A	Humidity, Min Temp, Wind Speed, PSL	Obs Precipitation	16:1
5A	Humidity, Min Temp, Wind Speed, PSL, Max Temp	Obs Precipitation	20:1
6A	Humidity, Wind Speed, PSL, Average Temp	Obs Precipitation	16:1
7A	Precipitation, Humidity, Min Temp, Wind Speed, PSL, Max Temp	Obs Precipitation	24:1

Table 6
Typical Model Architecture (Model 7A – IPSL_CMA5).

ANN Specifications	
Model Architecture	24:2:2:1
Number of Epochs	100
Transfer and Activation Functions	Logsig, logsig, purelin
Normalization Range	0,1
SVR Specifications	
Algorithm	Reg Optimizer: Reg SMO Improved
Kernel	The polynomial kernel
MGGP Specifications	
Population size	500
Max. generations	400
Generations elapsed	400
Input variables	24
Training instances	975
Tournament size	13
Elite fraction	0.5
Lexicographic selection pressure	On
Probability of pareto tournament	0.4
Max. genes	4
Max. tree depth	4
Max. total nodes	Inf
ERC probability	0.1
Crossover probability	0.78
Mutation probabilities	High level 0.2, Low level 0.8 0.2 Subtree 0.9, Input 0.05, Perturb ERC 0.05
Complexity measure	Expressional
Function set	TIMES MINUS PLUS RDIVIDE SQUARE MULT3 ADD3 SQRT CUBE

Table 7
Typical Phase 2 IPSL_CMA5 results (all values are in mm).

Model	Error Measure	DBS	ANN	SVR	MT	MGGP
1A	RMSE(mm)	125.23	70.41	83.84	72.52	75.89
	MAE(mm)	74.10	42.71	54.39	43.6	46.20
	CE	-0.11	0.64	0.5	0.63	0.59
	PEP(%)	15.09	-52.86	-72.23	-48.02	-64.94
2A	RMSE(mm)	125.23	65.99	82.8	70.43	73.84
	MAE(mm)	74.10	40.1	53.6	41.56	45.63
	CE	-0.11	0.69	0.51	0.65	0.61
	PEP(%)	15.09	-45.62	-72.36	-42.09	-53.43
3A	RMSE(mm)	125.23	63.95	73.9	71.83	70.28
	MAE(mm)	74.10	38.61	48.45	42.74	43.76
	CE	-0.11	0.71	0.61	0.63	0.65
	PEP(%)	15.09	-25.3	-59.12	-39.79	-45.32
4A	RMSE(mm)	125.23	64.34	71.34	62.64	67.83
	MAE(mm)	74.10	39.54	45.82	37.95	41.23
	CE	-0.11	0.71	0.64	0.72	0.67
	PEP(%)	15.09	-16.92	-56.51	-37.56	-51.67
5A	RMSE(mm)	125.23	61.3	72.01	71.28	101.53
	MAE(mm)	74.10	36.79	44.91	42.53	70.42
	CE	-0.11	0.73	0.63	0.64	0.27
	PEP(%)	15.09	2.41	-53.04	-40.91	-64.32
6A	RMSE(mm)	125.23	64.38	71.84	71.33	66.74
	MAE(mm)	74.10	38.88	45.28	42.98	40.52
	CE	-0.11	0.71	0.63	0.638	0.68
	PEP(%)	15.09	2.36	-54.94	-40.51	-43.57
7A	RMSE(mm)	125.23	60.63	71.4	63.4	67.84
	MAE(mm)	74.10	36.93	44.56	38.09	43.86
	CE	-0.11	0.74	0.64	0.71	0.67
	PEP(%)	15.09	-3.61	-55.28	-33.18	-48.79

The scatter plots for IPSL_CMA5 in Figure 3 show similar results which endorse the superiority of ANN models over the other soft computing models at least in this study. SVR and MGGP scatters are fairly distributed, with a propensity towards over-prediction of values. The scatter for MT is slightly better than for SVR and MGGP, but there are still a number of over-predicted

values. When compared to the other three approaches, the ANN scatter plot comes out on top as it is reasonably balanced. After examining the statistical error measurements and scatter plots, it can be concluded that the performance of ANN models beats that of SVR, MT, and MGGP models. In the case of IPSL_CMA5, the performance of MT appears to be the next in line, followed by MGGP and then SVR. ANN seems to have a better performance as compared to the other soft computing tools. This may be a result of the extra degrees of freedom that were introduced in the neural network architecture by adding the second hidden layer and making it a 4 layered network which may have improved its learning capabilities. All the soft tools are beating DBS comprehensively for the downscaling process by a considerable margin. Similar observations are mirrored in the result table of the remaining 4 GCMs (not shown here). This may be due to the inherent quality of soft computing techniques in that learning the physics of the process being modeled and providing results that are fitting for the parameter under consideration. Since ANN outperformed all other methods, only the best ANN models (in each GCM) will be studied in this paper further.

A rigorous decade-by-decade RMSE analysis was also undertaken for the best model of each GCM to ensure accuracy.

This task was undertaken to ensure the stability of the ANN models and to prevent them from oscillating toward extreme values. Table 8 illustrates that the RMSE value for each decade across all models remained within a consistent range, indicating the absence of model oscillation. Additionally, this exercise proved valuable for evaluating the performance of the GCMs. Table 8 highlights that IPSL_CMA5 demonstrated the highest effectiveness, followed by CNRM_CM5, with NorESMI_1 ranking third. As far as time required for model training is concerned, ANN, MT and SVR required few seconds while MGGP took 15 mins.

Table 8
Decade wise RMSE analysis (mm).

RMSE per 10 years	IPSL_CMA5	HadGEM2_CC	CNRM_CM5	NorESM1_1	GFDL_CM3
1901-1910	62.9	64.26	52.17	56.62	71.47
1911-1920	73.41	73.07	69.84	64.64	79.19
1921-1930	65.15	58.79	53.21	56.86	71.93
1931-1940	57.75	81.44	59.01	59.91	67.08
1941-1950	53.4	66.01	58.51	68.27	79.85
1951-1960	46.7	68.7	57.28	60.92	77.92
1961-1970	46.21	61.03	60.67	55.25	70.71
1971-1980	60.44	68.1	74.34	72.19	88.62
1981-1990	55.84	71.67	55.88	53.07	71.4
1991-2000	56.04	74.05	60.34	65.9	74.31
2001-2010	57.77	133.27	63.92	63.52	91.89
2011-2016	56.24	81.00	70.37	70.37	86.42
Average	57.65	75.12	61.29	62.29	77.57

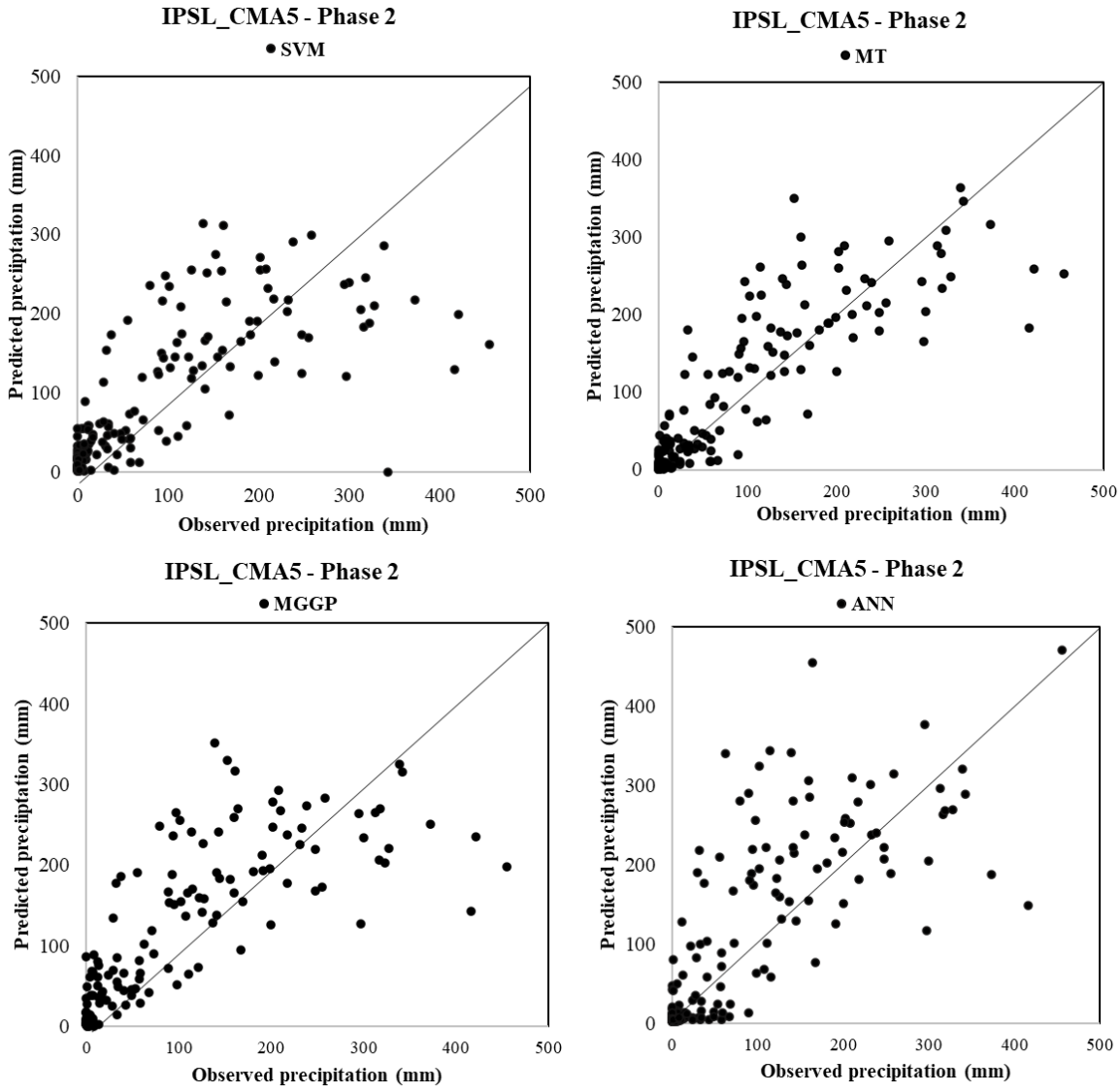

Fig. 3. (a to d) Scatter plots (IPSL_CMA5 - Phase2).

6. Precipitation trends in the future

As per the RMSE analysis, the top-performing models, namely 7A (IPSL_CMA5), 7C (CNRM_CM5), and 6B (NorESMI_1), were selected to forecast future rainfall in Pune for the next 30 years (2021–2050). Initially, the causative parameters used for training each of these models were extracted from the four grid points surrounding Pune for the RCP 4.5 condition spanning from 2021 to 2050, following the approach suggested by previous researchers such as [15] for India. Subsequently, these parameters were inputted into the pre-trained ANN models, with the output from the unseen models representing the monthly rainfall in Pune for the aforementioned period.

Tables 9 and 10 provide insights into the absolute and percentage changes anticipated in the future, respectively. These projections indicate notable shifts in future rainfall trends, suggesting

that the existing monsoon season in Pune may undergo significant alterations. Illustrated in Figure 4, the average rainfall is forecasted to witness substantial increases, ranging from 200–500% for January, 200–300% for February and March, and 100–150% for April and December. In contrast, rainfall appears to be decreasing by 20-30% between June and September, as observed in Figure 5. As a result, rain events may occur throughout the year rather than just from June to September. This indicates that Pune may soon be facing rainfall all year long.

Table 9
Comparison of Monthly average rainfall for past and future (2021-2050) years (mm).

Year/GCM	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1901-1930	2.83	1.58	1.99	10.01	22.83	146.86	271.55	178.69	146.81	57.22	28.65	4
1931-1960	2.19	0.42	2.71	11.47	28.18	164.9	377.82	244.38	183.51	99.82	37.84	4.96
1961-1990	0.69	0.78	1.88	6.12	27.33	169	354.96	257.87	173.69	73.21	23.51	7.16
1991-2016	1.03	7.15	6.23	5.32	16.7	189.1	239.54	182.93	141.93	94.86	23.11	6.53
Model 7A	7.17	8.1	11.52	21.41	64.67	184.1	234.13	214.43	149.87	112.39	43.71	14.52
Model 7C	11.23	8.25	11.43	17.95	117.99	238.84	232.64	204.07	149.93	89.64	46.17	17.28
Model 6B	7.44	9.86	11.45	11.73	40.38	202.76	217.15	195.68	175.99	94.23	31.51	9.29

Table 10
Percentage increase in future rainfall (2021-2050) (%).

Model/Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Model 7A	325.26	226	259.53	160.05	172.18	9.93	-24.7	-0.71	-7.19	38.27	54.58	156.38
Model 7C	565.97	231.97	256.86	118.03	396.57	42.62	-25.18	-5.5	-7.15	10.29	63.29	205.11
Model 6B	341.06	296.94	257.38	42.44	69.95	21.07	-30.17	-9.39	8.98	15.93	11.43	64.11

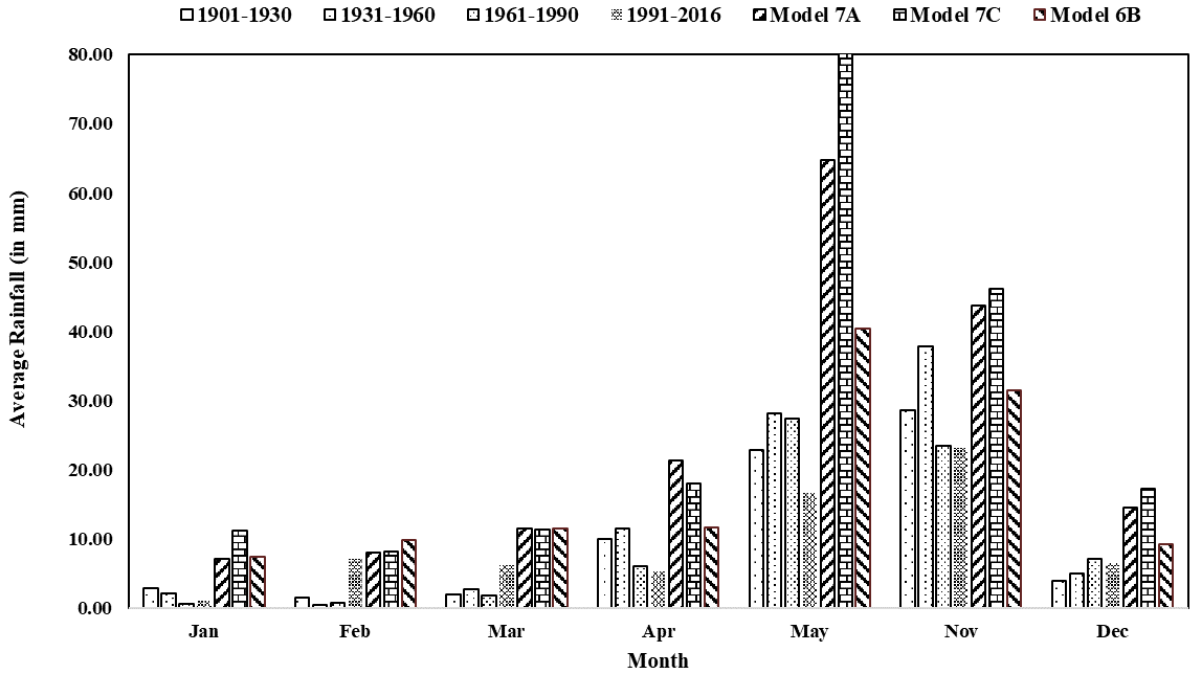

Fig. 4. Changes in average rainfall for Non-Monsoon months (2021-2050).

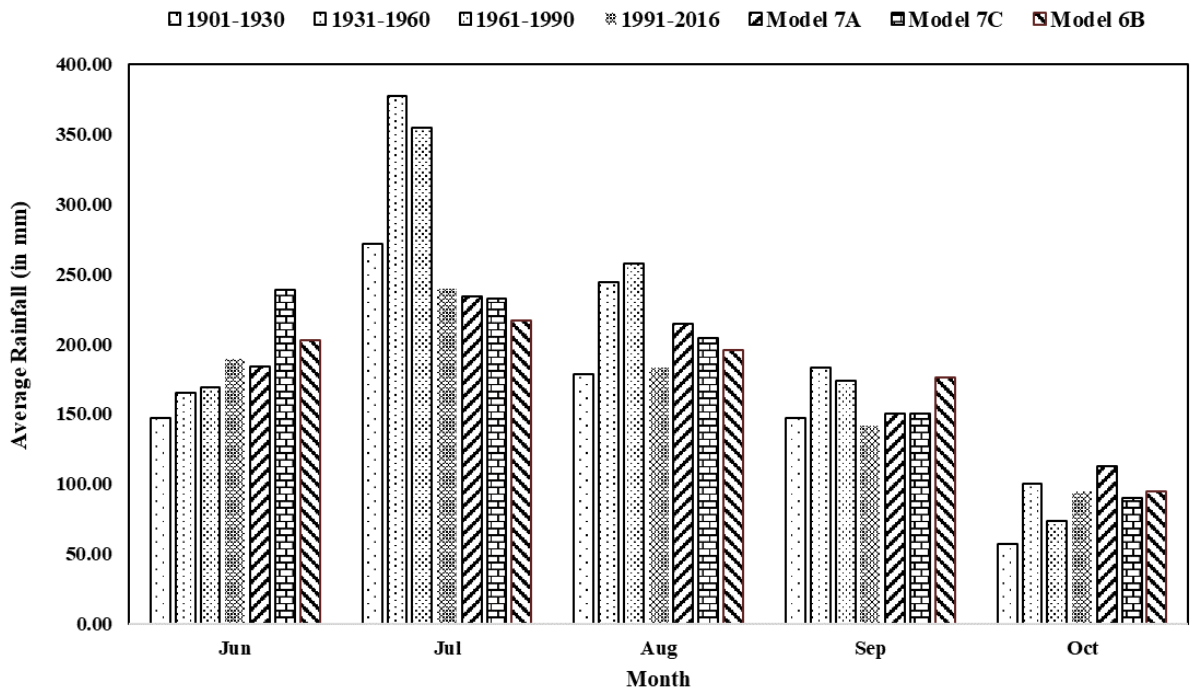

Fig. 5. Changes in average rainfall for Monsoon months (2021-2050).

7. Concluding remarks

The current study is an in-depth examination of the use of soft computing tools to downscale precipitation records extracted from various Global Circulation Models for the city of Pune over the next 30 years (2021-2050). There was no such study done the authors knowledge for the city of Pune which is the eighth largest city of India.

The primary innovation of this study lies in its methodology, which leverages the physical parameters obtained from GCMs at four adjacent grid points as input for training the ANN, SVR, MT, and MGGP models.

This approach enables the downscaling of parameters at the desired location (Pune) without the need for regridding the data, standardization, or bias correction over raw GCM data. Consequently, the output corresponds to the observed values at the desired location. Additionally, this study advocates for the utilization of Average Mutual Information (AMI), which is particularly well-suited for capturing nonlinear relationships, such as those found in rainfall data. The work is carried out in two phases with and without using downscaled values of causative parameters for estimating the rainfall. Following conclusions can be drawn from the above study

1. The results of phases 1 and 2 revealed that soft computing tools were far more accurate than DBS. Soft computing tools have an inherent ability to understand the physics behind the parameter under study, giving them an advantage over other techniques. ANN was the most accurate at downscaling the rainfall, followed by MT, then MGGP, and finally SVR.

2. Nine different training algorithms were tried and tested while modeling the neural networks, with LM being the most successful.
3. Initial Artificial Neural Networks (ANNs) with a single hidden layer failed to achieve the desired results. Consequently, the architecture was updated to include two hidden layers, resulting in a four-layered ANN, which yielded more satisfactory outcomes.
4. The results obtained from the ANN models did not exhibit oscillation, a finding corroborated by a detailed decade-wise RMSE analysis.
5. The current work also demonstrates that ANNs can accurately model precipitation even when some causative parameters are not provided as inputs during training. As a result, it validates neural network's efficiency even when there is missing data.
6. The present study sheds light on the potential reasons why the traditional monsoon season may not persist in the coming years. A significant increase in rainfall is projected for non-monsoon months, with a range of 200–500% for January, 200-300% for February and March, and 100-150% for April and December. Conversely, a decrease in rainfall ranging from 20-30% is anticipated between June and September. While numerous studies have explored precipitation trends across India, it is essential to model precipitation as a location-specific parameter accurately. Additionally, previous studies have not provided a detailed forecast of future monthly precipitation trends.
7. A shift in the monsoon season presents an opportunity to reconsider policies related to agricultural and water management activities. This aspect adds practical relevance to the current study, as it offers insights that could inform decision-making in these crucial sectors.

8. Declarations

8.1. Statement on data availability

Upon reasonable request, all data, models, or code supporting the findings of this study can be obtained from the corresponding author.

8.2. Financial disclosure

The authors declare no relevant financial or non-financial interests.

8.3. Contribution of authors

Both authors contributed to the conception and design of the study. Material preparation, data collection, and analysis were conducted by both authors.

8.4. Availability of code

Upon reasonable request, all models or code supporting the findings of this study can be obtained from the corresponding author.

8.5. Ethics approval

The authors declare no conflicts of interest concerning the content of this article.

8.6. Participation consent

Informed consent was obtained from all individual authors included in the study.

8.7. Publication consent

Both authors have agreed to submit the technical paper to the journal.

References

[1] Todmal RS. Future Climate Change Scenario over Maharashtra, Western India: Implications of the Regional Climate Model (REMO-2009) for the Understanding of Agricultural Vulnerability. Pure Appl Geophys 2021;178:155–68. https://doi.org/10.1007/s00024-020-02642-6.

[2] Dias, A., Dhawde, R., Surve, N., Weinberg, A., Birdi, T. and Mistry N. Impact of climate changes on water availability and quality in the state of maharashtra in western India. Asian Jr Microbiol Biotech Env Sc 2015;17:1071–81.

[3] Deepthi B, Sivakumar B. General circulation models for rainfall simulations: Performance assessment using complex networks. Atmos Res 2022;278:106333. https://doi.org/10.1016/j.atmosres.2022.106333.

[4] Change IC. The physical science basis. Cambridge Univ Press Cambridge, United Kingdom New York, NY, USA 2013.

[5] Ghosh S, Mujumdar PP. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. Adv Water Resour 2008;31:132–46. https://doi.org/10.1016/j.advwatres.2007.07.005.

[6] Kripalani RH, Kulkarni A, Sabade SS, Khandekar ML. Indian Monsoon Variability in a Global Warming Scenario. Nat Hazards 2003;29:189–206. https://doi.org/10.1023/A:1023695326825.

[7] Ueda H, Iwai A, Kuwako K, Hori ME. Impact of anthropogenic forcing on the Asian summer monsoon as simulated by eight GCMs. Geophys Res Lett 2006;33. https://doi.org/10.1029/2005GL025336.

[8] Stowasser M, Annamalai H, Hafner J. Response of the South Asian Summer Monsoon to Global Warming: Mean and Synoptic Systems. J Clim 2009;22:1014–36. https://doi.org/https://doi.org/10.1175/2008JCLI2218.1.

[9] Sabade SS, Kulkarni A, Kripalani RH. Projected changes in South Asian summer monsoon by multi-model global warming experiments. Theor Appl Climatol 2011;103:543–65. https://doi.org/10.1007/s00704-010-0296-5.

[10] Krishnan R, Sabin TP, Ayantika DC, Kitoh A, Sugi M, Murakami H, et al. Will the South Asian monsoon overturning circulation stabilize any further? Clim Dyn 2013;40:187–211. https://doi.org/10.1007/s00382-012-1317-0.

[11] Lal M, Meehl GA, Arblaster JM. Simulation of Indian summer monsoon rainfall and its intraseasonal variability in the NCAR climate system model. Reg Environ Chang 2000;1:163–79. https://doi.org/10.1007/s101130000017.

[12] Meehl GA, Arblaster JM. Mechanisms for projected future changes in south Asian monsoon precipitation. Clim Dyn 2003;21:659–75. https://doi.org/10.1007/s00382-003-0343-3.

[13] Kumar KR, Sahai AK, Kumar KK, Patwardhan SK, Mishra PK, Revadekar J V, et al. High-resolution climate change scenarios for India for the 21st century. Curr Sci 2006;90:334–45.

[14] Mohammadkhani MR, Shishegaran A, Shokrollahi B. Forecasting probable maximum precipitation using innovative algorithm to estimate atmosphere precipitable water vapor. Math Model Eng 2019;5:90–6. https://doi.org/10.21595/mme.2019.20935.

[15] Rana A, Foster K, Bosshard T, Olsson J, Bengtsson L. Impact of climate change on rainfall over Mumbai using Distribution-based Scaling of Global Climate Model projections. J Hydrol Reg Stud 2014;1:107–28. https://doi.org/10.1016/j.ejrh.2014.06.005.

[16] Raje D, Mujumdar PP. A comparison of three methods for downscaling daily precipitation in the Punjab region. Hydrol Process 2011;25:3575–89. https://doi.org/10.1002/hyp.8083.

[17] Singh K, Sihag P, Kumar R. Downscaling monthly rainfall using support vector regression and M5P model tree. 5th Int. Conf. "Architecture, Civ. Environ. Eng. New Delhi, India"(ACEE–2014), 2014.

[18] Vu MT, Aribarg T, Supratid S, Raghavan S V, Liong S-Y. Statistical downscaling rainfall using artificial neural network: significantly wetter Bangkok? Theor Appl Climatol 2016;126:453–67. https://doi.org/10.1007/s00704-015-1580-1.

[19] Xu R, Chen N, Chen Y, Chen Z. Downscaling and Projection of Multi-CMIP5 Precipitation Using Machine Learning Methods in the Upper Han River Basin. Adv Meteorol 2020;2020:1–17. https://doi.org/10.1155/2020/8680436.

[20] Wu Y, Zhang Z, Crabbe MJC, Chandra Das L. Statistical Learning-Based Spatial Downscaling Models for Precipitation Distribution. Adv Meteorol 2022;2022:1–12. https://doi.org/10.1155/2022/3140872.

[21] Searson DP, Leahy DE, Willis MJ. GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. Proc. Int. multiconference Eng. Comput. Sci., vol. 1, Citeseer; 2010, p. 77–80.

[22] Artificial Neural Networks in Hydrology. I: Preliminary Concepts. J Hydrol Eng 2000;5:115–23. https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(115).

[23] Khan SR, Noor S. Short Term Load Forecasting using SVM based PUK kernel. 2020 3rd Int. Conf. Comput. Math. Eng. Technol., IEEE; 2020, p. 1–9. https://doi.org/10.1109/iCoMET48670.2020.9074068.

[24] Solomatine DP, Xue Y. M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. J Hydrol Eng 2004;9:491–501. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491).

[25] Dawson CW, Abrahart RJ, See LM. HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ Model Softw 2007;22:1034–52. https://doi.org/https://doi.org/10.1016/j.envsoft.2006.06.008.

[26] Londhe SN, Deo MC. Wave tranquility studies using neural networks. Mar Struct 2003;16:419–36. https://doi.org/https://doi.org/10.1016/j.marstruc.2003.09.001.

[27] Demuth, H., Beale, M., Hagen M. Neural Network Toolbox User's guide. The Mathworks INC; 1998.

[28] Waikato Environment for Knowledge Analysis (WEKA) Software version 3. 8. 6. https://www.cs.waikato.ac.nz/~ml/weka n.d.

[29] Okan F, Umut O. Statistical Downscaling of Monthly Precipitation Using NCEP/NCAR Reanalysis Data for Tahtali River Basin in Turkey. J Hydrol Eng 2011;16:157–64. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000300.

[30] Parisa S, Omid B-H, Atiyeh B, A. LH. Runoff Projection under Climate Change Conditions with Data-Mining Methods. J Irrig Drain Eng 2017;143:4017026. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001205.

[31] Abdullahi J, Elkiran G. Prediction of the future impact of climate change on reference evapotranspiration in Cyprus using artificial neural network. Procedia Comput Sci 2017;120:276–83. https://doi.org/https://doi.org/10.1016/j.procs.2017.11.239.

[32] Rabezanahary Tanteliniaina MF, Rahaman MH, Zhai J. Assessment of the Future Impact of Climate Change on the Hydrology of the Mangoky River, Madagascar Using ANN and SWAT. Water 2021;13. https://doi.org/10.3390/w13091239.

[33] Londhe SN, Shah S. A novel approach for knowledge extraction from Artificial Neural Networks. ISH J Hydraul Eng 2019;25:269–81. https://doi.org/10.1080/09715010.2017.1409667.

[34] Gowariker V, Thapliyal V, Sarker RP, Mandal GS, Sikka DR. Parametric and power regression models: New approach to long range forecasting of monsoon rainfall in India. Mausam 1989;40.